JOURNAL OF BACTERIOLOGY, Nov. 2010, p. 6045–6055 Vol. 192, No. 22
0021-9193/10/$12.00 doi:10.1128/JB.00277-10

Predicting Plasmid Promiscuity Based on Genomic Signature▽‡

Haruo Suzuki,† Hirokazu Yano, Celeste J. Brown, and Eva M. Top*

Department of Biological Sciences, Initiative for Bioinformatics and Evolutionary Studies (IBEST), University of Idaho, Moscow, Idaho 83844-3051

Received 11 March 2010/Accepted 3 September 2010

Despite the important contribution of self-transmissible plasmids to bacterial evolution, little is understood about the range of hosts in which these plasmids have evolved. Our goal was to infer this so-called evolutionary host range. The nucleotide composition, or genomic signature, of plasmids is often similar to that of the chromosome of their current host, suggesting that plasmids acquire their hosts' signature over time. Therefore, we examined whether the evolutionary host range of plasmids could be inferred by comparing their trinucleotide composition to that of all completely sequenced bacterial chromosomes. The diversity of candidate hosts was determined using taxonomic classification and genetic distance. The method was first tested using plasmids from six incompatibility (Inc) groups whose host ranges are generally thought to be narrow (IncF, IncH, and IncI) or broad (IncN, IncP, and IncW) and then applied to other plasmid groups. The evolutionary host range was found to be broad for IncP plasmids, narrow for IncF and IncI plasmids, and intermediate for IncH and IncN plasmids, which corresponds with their known host range. The IncW plasmids as well as several plasmids from the IncA/C, IncP, IncQ, IncU, and PromA groups have signatures that were not similar to any of the chromosomal signatures, raising the hypothesis that these plasmids have not been ameliorated in any host due to their promiscuous nature. The inferred evolutionary host range of IncA/C, IncP-9, and IncL/M plasmids requires further investigation. In this era of high-throughput sequencing, this genomic signature method is a useful tool for predicting the host range of novel mobile elements.

Comparative genomics has clearly shown that bacterial evolution occurs not only through genetic changes that are vertically inherited but also by extensive horizontal gene transfer between closely and distantly related bacteria (9). Mobile genetic elements such as plasmids and phages serve as important agents of horizontal gene transfer that can exchange genetic material between chromosomes (26). Plasmids also play a critical role in rapid bacterial adaptation to local environmental changes, as best exemplified by the alarmingly rapid spread of plasmid-encoded multidrug resistance in human pathogens (44, 66). In spite of this, very little is understood about the range of bacterial hosts in which these plasmids may have resided and evolved in natural or clinical environments over time, i.e., their potential "evolutionary host range." Understanding the evolutionary history of virulence, catabolic, and other plasmids may help us to reconstruct the plasmid transfer network among microorganisms and track the pathways of gene dissemination.

A plasmid's host range can be defined in different ways, but it is typically understood as the range of hosts in which a plasmid can replicate (replication host range, or from here on simply called "host range"). This host range is often narrower than the range of hosts to which the plasmid can transfer by conjugation (transfer host range) (32, 72) but wider than the range in which it can be stably maintained (long-term host range) (16). The host range of a plasmid is often determined by mating assays, wherein that plasmid is transferred into a set of recipient strains followed by selection for transconjugant clones that can express one of the traits encoded by the plasmid (40, 47). Ideally, the physical presence of the plasmids is then verified to confirm independent replication. Sometimes the host range is also inferred from the observed natural range of hosts in which a plasmid is found in various habitats (24, 72). The plasmid host range is known to be highly variable among plasmids, and the terms "narrow host range" and "broad host range" are used as qualitative indicators (18, 49, 62). For example, it has been generally considered that incompatibility (Inc) groups IncF, IncH, and IncI contain self-transmissible narrow-host-range plasmids, while IncN, IncP, and IncW plasmids transfer and replicate in a broad range of hosts (13, 49, 62). This oldest system of plasmid classification into Inc groups is based on the inability of plasmids from the same group to be maintained in the same host due to similarity in replication or partitioning systems (11, 53). We note that IncP plasmids are also called IncP-1 in the *Pseudomonas* classification system, but they are here referred to as IncP. The entire range of hosts, including ancestral forms and extant bacteria, in which a plasmid has replicated at some point during its evolutionary history is of course unknown but expected to be narrower than its replication range. Here, we designate this range the "evolutionary host range."

To understand the contributions of plasmids to horizontal gene transfer and bacterial evolution, it is not sufficient to know the hosts in which plasmids can potentially replicate and be maintained when tested in the laboratory or the field. While very valid, such experiments (13, 17, 40, 47, 56, 72) do not allow

* Corresponding author. Mailing address: Department of Biological Sciences, P.O. Box 443051, University of Idaho, Moscow, ID 83844-3051. Phone: (208) 885-5015. Fax: (208) 885-7905. E-mail: evatop@uidaho.edu.

† Present address: Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853.

‡ Supplemental material for this article may be found at http://jb.asm.org/.

▽ Published ahead of print on 27 September 2010.

us to evaluate which plasmids have in fact spread among the widest range of hosts in the past and therefore contributed most so far to horizontal gene transfer across distantly related bacteria. We also need to gain insight into the range of hosts in which they have actually resided over evolutionary time—their evolutionary host range. This insight into the evolutionary history of plasmids will also shed light on the reservoirs of the many unwanted drug resistance and virulence plasmids (65). Previous studies have shown that the dinucleotide composition (2-mer genomic signatures) of plasmids tend to be similar to those of the chromosomes of their known host, suggesting that the plasmids acquire the host's genomic signature (7, 67). It has previously been suggested that host-specific mutational biases homogenize the nucleotide compositions of genetic elements that are being replicated in the same host (plasmids, phages, and DNA fragments inserted in the chromosome); this phenomenon has been designated "genome amelioration" (7, 43). In addition, due to the potential DNA exchange between chromosomes and plasmids by recombination and transposition (8, 42), acquisition of large sections of chromosomal DNA by plasmids may also result in similar signatures between plasmids and their evolutionary hosts. It thus follows that a similar genomic signature between a plasmid and a host's chromosome may indicate residence of the plasmid in that or a closely related host during its evolutionary history. Therefore, it should be possible to infer the evolutionary host range for plasmids whose genome sequences have been determined, based on the similarity in genomic signature with that of completely sequenced bacterial chromosomes.

The goal of this study was to infer the evolutionary host range of various plasmids based on their genomic signatures. Specifically, we postulate (i) that known broad-host-range plasmids from *Proteobacteria* have evolved in a wider range of hosts than narrow-host-range plasmids and (ii) that our genomic signature approach can be used to assess the promiscuity of sequenced but uncharacterized plasmids and other mobile elements. To develop our approach, we chose self-transmissible plasmids belonging to six incompatibility groups, whose host ranges have been studied intensively and are thought to be narrow (IncF, IncH, and IncI) or broad (IncN, IncP, and IncW). To propose candidate evolutionary hosts of these plasmids, we compared the genomic signature of each plasmid with those of 817 chromosomes of prokaryotes for which complete sequences were available. Our results suggest that the evolutionary host range is broad for IncP plasmids, narrow for IncF and IncI plasmids, and intermediate for IncH and IncN plasmids. The lack of hosts with signatures similar to the IncW plasmids raises the hypothesis that they have not been ameliorated for any host due to their promiscuity. We then used the same method to infer the evolutionary host range of additional plasmid groups, such as IncA/C (also called IncP-3), IncL/M, IncP-9, IncQ (IncP-4), IncU, and PromA and plasmids Ri and Ti from *Agrobacterium* sp. (designated Ri/Ti). The similarities and discrepancies between our findings and previous knowledge on plasmid host range are discussed.

MATERIALS AND METHODS

Software. Genome analyses were conducted using G-language genome analysis environment version 1.8.10 (2), available at http://www.g-language.org/. Statistical tests and graphics were implemented using R version 2.10.1 (57), available at http://www.r-project.org/.

Genome sequences. Completely sequenced genomes of plasmids and bacterial chromosomes were downloaded in GenBank format (5) from the National Center for Biotechnology Information (NCBI) site (http://www.ncbi.nlm.nih.gov/) in July 2009. In cases in which the bacterial strain has multiple chromosomes, only the largest chromosome was used for the analysis. For the complete listings of the 1,945 plasmids and 817 chromosomes (56 *Archaea* and 761 *Bacteria*) used in this study, see Tables S1 and S2 in the supplemental material. A few plasmids were (re)named by us: pTi (GenBank accession number NC_002377), pTi-C58 (NC_003065), pAMMD_1 (NC_008385), pKP9 (NC_011383), and pKP12 (NC_011385).

Identifying plasmid incompatibility groups. We performed a protein homology search to determine which of the 1,945 completely sequenced plasmids likely belong to the following 6 incompatibility (Inc) groups with well-known host ranges: IncF, IncH, IncI, IncN, IncP, and IncW. Homologous proteins were inferred using BLASTP (1) with an E-value cutoff of $1e^{-5}$. We used 14 reference plasmids previously classified by traditional incompatibility typing into the IncF, IncH, IncI, IncN, IncP, or IncW group: F and R100 belong to IncF (subgroups IncFI and IncFII, respectively); R27 and R478 to IncH (subgroup IncHI); ColIb-P9 and R64 to IncI; R46 to IncN; RK2, R751, pJP4, pQKH54, and pKJK5 to IncP; and R388 and R7K to IncW (4, 10, 12, 14, 15, 20, 33, 35, 37, 39, 51, 71, 77). From the 1,945 plasmids, we retrieved plasmids that met both of the following two criteria: (i) they encode proteins that are homologs of more than half of all proteins encoded by any of the reference plasmids, and (ii) they encode replication initiation (Rep) proteins that are homologs to those of the reference plasmids: RepB and RepE from IncFI plasmid F, RepA1 and RepA4 from IncFII plasmid R100, RepHIA from the IncH plasmids, RepZ from the IncI plasmids, RepA from the IncN plasmid, TrfA from the IncP plasmids, and RepA from the IncW plasmids (3, 19, 23, 24, 29, 38, 48, 63, 79). The combination of two criteria allowed the discrimination of those plasmid genomes that share many proteins and encode the same replication machineries as the reference genomes (thus satisfying both criteria) from those that just share many proteins but not the Rep proteins, or only Rep but very few other proteins. This approach was used because given the mosaic nature of plasmids, some may very well share many proteins with a given reference plasmid but have a very different replicon and thus would almost certainly belong to a different Inc group and have a different host range. We chose the incompatibility group classification, rather than a more recently proposed classification system based on the transfer-related relaxase protein (28), because incompatibility grouping is tightly correlated with replication systems, and the host range is typically limited by a plasmid's ability to replicate, not its ability to transfer.

The same approach was used to retrieve plasmids belonging to other well-known groups that contain multiple completely sequenced plasmids: IncA/C (also called IncP-3), IncL/M, IncP-9, IncQ (IncP-4), and IncU. The replication initiation proteins used to retrieve the plasmids and their corresponding reference plasmids were RepA from the IncA/C plasmid pRA1 (34, 45), RepA from the IncL/M plasmid pCTX-M3 (52), Rep from the IncP-9 plasmid NAH7 and pWW0 (60, 61), RepC from the IncQ plasmid RSF1010 (50, 59), and RepB from the IncU plasmid pRA3 (41). The PromA (74) and Ri/Ti (68) plasmids used here were retrieved from the literature and include those representatives whose genome sequences were available in GenBank in July 2009.

To visualize and confirm our homology-based assignment of plasmids to these six Inc groups, we performed cluster analysis of these plasmids based on their patterns of gene content (6, 64). We performed all-against-all protein sequence comparisons (BLASTP with an E-value cutoff of $1e^{-5}$) followed by Markov clustering with an inflation factor of 2.0 (75) for constructing a group of homologous proteins (here referred to as a "protein family"). For the obtained gene content table, see Table S3 in the supplemental material. The dissimilarity in gene content patterns (binary data for the presence or absence of each protein family) between two plasmids was measured by the Jaccard distance (one minus the Jaccard coefficient), and the distance matrix was subject to hierarchical cluster analysis (unweighted pair group method with arithmetic mean [UPGMA]) (25).

Calculating the genomic signature of a DNA sequence. The trinucleotide composition or 3-mer genomic signature of a DNA sequence (plasmid or chromosome) was represented by a vector, which consists of 64 trinucleotide relative abundance values. The trinucleotide relative abundance value (x_{ijk}) is defined as the observed trinucleotide frequency divided by the expected trinucleotide frequency, which is the product of the component mononucleotide frequencies:

$$x_{ijk} = \frac{f_{ijk}}{f_i f_j f_k}$$

TABLE 1. Diversity of candidate evolutionary hosts of 92 plasmids

Name	Group[a]	Known host(s)[b]	No. of taxa at indicated rank[c]						*D*mean[d]	
			Class	Order	Family	Genus	Species	Strain	16S rRNA	3-mer
IncF, IncH, IncI, IncN, IncP, and IncW plasmids										
F	IncF*	*Escherichia coli*	1	1	1	1	1	9	0.008	0.006
p1658/97	IncF	*Escherichia coli*	1	1	1	3	7	42	0.016	0.033
pC15-1a	IncF	*Escherichia coli*	1	1	1	3	5	26	0.012	0.020
pIP1206	IncF	*Escherichia coli*	1	1	1	3	7	43	0.017	0.034
pO26I	IncF	*Escherichia coli*	1	1	1	1	1	12	0.008	0.006
pO86A1	IncF	*Escherichia coli*	1	1	1	2	2	14	0.008	0.006
pSMS35_130	IncF	*Escherichia coli*	1	1	1	3	6	34	0.016	0.031
R100	IncF*	*Shigella flexneri*	1	1	1	3	6	34	0.015	0.031
pAPEC-O1-R	IncH	*Escherichia coli*	1	1	1	11	15	54	0.029	0.054
pEC-IMP	IncH	*Enterobacter cloacae*	1	1	1	13	19	65	0.035	0.064
pHCM1	IncH	*Salmonella enterica*	1	2	2	13	19	65	0.037	0.063
pK29	IncH	*Klebsiella pneumoniae*	1	1	1	11	16	60	0.032	0.060
pMAK1	IncH	*Salmonella enterica*	1	2	2	12	18	64	0.036	0.063
R27	IncH*	*Salmonella enterica*	1	2	2	6	10	47	0.034	0.054
R478	IncH*	*Serratia marcescens*	1	1	1	8	12	47	0.030	0.053
ColIb-P9	IncI*	*Shigella sonnei*	1	1	1	2	2	13	0.008	0.006
pCVM29188_101	IncI	*Salmonella enterica*	1	1	1	2	2	17	0.008	0.007
pO113	IncI	*Escherichia coli*	1	1	1	1	1	10	0.009	0.006
pO26-Vir	IncI	*Escherichia coli*	1	1	1	1	1	10	0.008	0.006
pSE11-1	IncI	*Escherichia coli*	1	1	1	2	2	16	0.008	0.007
pSL476_91	IncI	*Salmonella enterica*	1	1	1	1	1	11	0.008	0.006
R64	IncI*	*Salmonella enterica*	1	1	1	3	6	28	0.013	0.024
pKP12	IncN	*Klebsiella pneumoniae*	1	1	1	13	18	61	0.032	0.059
pKP9	IncN	*Klebsiella pneumoniae*	1	1	1	12	14	48	0.027	0.050
pKP96	IncN	*Klebsiella pneumoniae*	1	1	1	11	11	28	0.031	0.056
pLEW517	IncN	*Escherichia coli*	1	1	1	12	13	40	0.029	0.054
pMAK2	IncN	*Salmonella enterica*	1	1	1	12	12	42	0.029	0.053
pMUR050	IncN	*Escherichia coli*	1	1	1	10	13	54	0.035	0.063
R46	IncN*	*Salmonella typhimurium*	1	1	1	13	17	56	0.028	0.052
pA1	IncP	*Sphingomonas* sp.	0	0	0	0	0	0	NA	NA
pA81	IncP	*Achromobacter xylosoxidans*	2	5	7	8	18	25	0.138	0.107
pADP-1	IncP	*Pseudomonas* sp.	2	5	7	12	24	35	0.138	0.109
pAMMD_1	IncP	*Burkholderia ambifaria*	1	1	1	1	1	1	NA	NA
pAOVO02	IncP	*Acidovorax* sp.	2	2	4	8	12	12	0.118	0.103
pB10	IncP	NA	2	2	4	5	9	10	0.126	0.097
pB3	IncP	NA	1	1	1	1	1	1	NA	NA
pB4	IncP	NA	2	4	6	7	11	13	0.132	0.102
pB8	IncP	NA	1	1	2	2	3	3	0.076	0.090
pBP136	IncP	*Bordetella pertussis*	1	1	1	1	1	1	NA	NA
pBS228	IncP	*Pseudomonas aeruginosa*	1	3	3	3	3	3	0.122	0.086
pCNB	IncP	*Comamonas* sp.	2	7	9	15	27	32	0.134	0.117
pEST4011	IncP	*Achromobacter denitrificans*	3	5	7	15	23	29	0.133	0.123
pIJB1	IncP	*Burkholderia cepacia*	2	4	6	12	20	27	0.126	0.110
pJP4	IncP*	*Ralstonia eutropha*	3	5	7	12	17	19	0.117	0.107
pKJK5	IncP*	NA	1	1	1	1	1	1	NA	NA
pQKH54	IncP*	NA	2	3	3	3	6	11	0.073	0.086
pTB11	IncP	NA	2	2	3	3	3	3	0.191	0.145
pTP6	IncP	NA	1	1	2	2	2	2	NA	NA
pUO1	IncP	*Delftia acidovorans*	2	4	6	11	17	21	0.131	0.107
R751	IncP*	*Enterobacter aerogenes*	0	0	0	0	0	0	NA	NA
RK2[e]	IncP*	*Klebsiella aerogenes* and *Pseudomonas aeruginosa*	1	1	1	1	2	2	NA	NA
pIE321	IncW	*Salmonella enterica*	0	0	0	0	0	0	NA	NA
pMAK3	IncW	*Salmonella enterica*	0	0	0	0	0	0	NA	NA
R388	IncW*	*Escherichia coli*	0	0	0	0	0	0	NA	NA
R7K	IncW*	*Providencia rettgeri*	0	0	0	0	0	0	NA	NA
IncA/C, IncL/M, IncP-9, IncQ, IncU, PromA, and Ri/Ti plasmids										
pAM04528	IncA/C	*Salmonella enterica*	2	2	3	3	4	4	0.167	0.113
pAR060302	IncA/C	*Escherichia coli*	2	4	5	5	6	6	0.167	0.122
pAsa4	IncA/C	*Aeromonas salmonicida*	2	3	3	3	4	4	0.161	0.098
peH4H	IncA/C	*Escherichia coli*	3	7	8	8	11	15	0.155	0.133
pIP1202	IncA/C	*Yersinia pestis*	2	3	4	4	5	5	0.171	0.111
pP91278	IncA/C	*Photobacterium damselae*	1	1	1	1	1	1	NA	NA
pP99-018	IncA/C	*Photobacterium damselae*	2	2	2	2	2	2	NA	NA
pRA1	IncA/C*	*Aeromonas hydrophila*	1	1	1	1	1	1	NA	NA
pSN254	IncA/C	*Salmonella enterica*	3	6	7	7	8	8	0.178	0.130
pYR1	IncA/C	*Yersinia ruckeri*	0	0	0	0	0	0	NA	NA
pCTX-M3	IncL/M*	*Citrobacter freundii*	1	1	1	12	17	59	0.030	0.057
pCTXM360	IncL/M	*Klebsiella pneumoniae*	1	1	1	4	7	37	0.018	0.035
pEL60	IncL/M	*Erwinia amylovora*	1	1	1	5	8	41	0.019	0.038
NAH7	IncP-9*	*Pseudomonas putida*	2	8	9	10	13	17	0.138	0.113
pDTG1	IncP-9	*Pseudomonas putida*	2	8	10	10	13	20	0.122	0.123
pNAH20	IncP-9	*Pseudomonas fluorescens*	2	8	10	11	14	21	0.125	0.123
pWW0	IncP-9*	*Pseudomonas putida*	2	5	5	5	12	20	0.091	0.104

Continued on following page

TABLE 1—*Continued*

Name	Group[a]	Known host(s)[b]	No. of taxa at indicated rank[c]						*D*mean[d]	
			Class	Order	Family	Genus	Species	Strain	16S rRNA	3-mer
pCCK1900	IncQ	*Pasteurella multocida*	1	1	1	1	1	1	NA	NA
pCHE-A	IncQ	*Enterobacter cloacae*	0	0	0	0	0	0	NA	NA
pDN1	IncQ	*Dichelobacter nodosus*	0	0	0	0	0	0	NA	NA
pIE1115	IncQ	NA	0	0	0	0	0	0	NA	NA
pIE1130	IncQ	NA	0	0	0	0	0	0	NA	NA
pMS260	IncQ	*Actinobacillus pleuropneumoniae*	0	0	0	0	0	0	NA	NA
RSF1010	IncQ*	*Escherichia coli*	0	0	0	0	0	0	NA	NA
pFBAOT6	IncU	*Aeromonas punctata*	1	1	1	1	1	1	NA	NA
pRA3	IncU*	*Aeromonas hydrophila*	0	0	0	0	0	0	NA	NA
pIPO2T	PromA	NA	0	0	0	0	0	0	NA	NA
pMRAD02	PromA	*Methylobacterium radiotolerans*	3	6	8	13	21	27	0.158	0.125
pSB102	PromA	NA	1	1	1	1	1	1	NA	NA
pTer331	PromA	*Collimonas fungivorans*	0	0	0	0	0	0	NA	NA
pRi1724	Ri/Ti	*Agrobacterium rhizogenes*	1	1	4	10	16	20	0.080	0.078
pRi2659	Ri/Ti	*Agrobacterium rhizogenes*	1	1	4	10	17	21	0.081	0.080
pTi	Ri/Ti	*Agrobacterium tumefaciens*	2	3	7	9	12	15	0.100	0.085
pTi-C58	Ri/Ti	*Agrobacterium tumefaciens*	1	2	6	12	15	19	0.097	0.091
pTi-SAKURA	Ri/Ti	*Agrobacterium tumefaciens*	1	2	6	11	14	17	0.096	0.088
pTiBo542	Ri/Ti	*Agrobacterium tumefaciens*	2	3	7	9	12	15	0.100	0.085
pTiS4	Ri/Ti	*Agrobacterium vitis*	1	2	6	12	18	21	0.083	0.078

[a] Group, incompatibility group based on experimental testing (*) or homology-based screening. See Table S4 in the supplemental material for the GenBank accession numbers of the plasmids.
[b] Known host, bacterial species name of the host in which the plasmid was found; NA, not available.
[c] All predicted hosts belong to one domain (*Bacteria*) and one phylum (*Proteobacteria*).
[d] *D*mean, mean distance between all pairs of candidate evolutionary host strains based on dissimilarity in 16S rRNA gene sequence and in 3-mer genomic signature. NA, not available because less than three strains were detected.
[e] The RK2 sequence is a composite genome sequence assembled from sequence information on presumably identical plasmids R18, R68, RK2, RP1, and RP4, which were found in different hosts, *Klebsiella aerogenes* and *Pseudomonas aeruginosa* (36, 54).

where f_i, f_j, and f_k denote the frequency of the mononucleotide i, j, and k, respectively ($i, j, k \in$ A, C, G, T) and f_{ijk} denotes the frequency of the trinucleotide ijk. These values combine counts from both strands of the sequence.

Measuring genomic signature difference between DNA sequences. The dissimilarity in 3-mer genomic signatures between two DNA sequences (plasmid-plasmid or chromosome-chromosome) was measured by average absolute difference (δ).

$$\delta = \frac{1}{64}\sum_i\sum_j\sum_k \left| x_{ijk} - y_{ijk} \right|$$

where x_{ijk} and y_{ijk} are the relative abundance values of the trinucleotide ijk for the sequences x and y, respectively, and the sum extends over all 64 trinucleotides. The distance matrix for plasmids was subject to multidimensional scaling (25).

Inferring evolutionary hosts of a plasmid based on genomic signature. We proposed candidate evolutionary hosts for each plasmid based on 3-mer genomic signature similarity between the plasmid and 817 bacterial chromosomes. The dissimilarity in genomic signature between each plasmid and a set of nonoverlapping 5-kb chromosomal segments from one bacterial strain was measured by the Mahalanobis distance, and the distances were converted to P values as described previously (67). High P values of close to 1 indicate small Mahalanobis distances and similar genomic signatures between a plasmid and chromosome; e.g., a P value of >0.6 indicates that the plasmid has a smaller Mahalanobis distance and thus is more similar to the average chromosome signature than >60% of the chromosomal segments. A bacterial strain was proposed as a candidate evolutionary host when a P value derived from Mahalanobis distance between the plasmid and chromosome was greater than 0.6.

Analyzing taxonomic range and diversity of evolutionary hosts. The taxonomic range of candidate evolutionary hosts for each plasmid was represented by the number of different taxonomic groups at all ranks; i.e., domain, phylum, class, order, family, genus, and species. The diversity of candidate evolutionary hosts for each plasmid was also quantified by a mean distance (*D*mean) between all pairs of the candidate evolutionary hosts (76). Distances were measured in two ways: on the basis of their dissimilarities in (i) 16S rRNA gene sequence and (ii) 3-mer genomic signature. The *D*mean value of a plasmid was calculated only when more than two candidate evolutionary hosts were available for that plasmid. Multiple alignments and distance measures of 16S rRNA sequences were implemented by MUSCLE (21). The 3-mer genomic signature consisted of a mean vector of trinucleotide relative abundance values calculated from the chromosomal segments, and the genomic signature dissimilarity between two chromosomes was measured by the average absolute difference (δ).

RESULTS

Retrieval of plasmids belonging to six incompatibility groups. To develop our genomic signature method for inferring the evolutionary host range of plasmids, we first retrieved the genomes of self-transmissible plasmids that belong to six incompatibility groups with well-studied host range characteristics. A homology-based screening procedure was applied to all 1,945 completely sequenced plasmids available in GenBank to retrieve plasmids that likely belong to the IncF, IncH, IncI, IncN, IncP, or IncW group. Fifty-nine plasmids were retained after this screening. From these, the following four plasmids were not included because they were regarded as duplicates: plasmid NR1 (NC_009133), plasmid EC-IMPQ (NC_012556), a transconjugant version of pLEW517 (NC_009131), and *Ralstonia eutropha* JMP134 plasmid 1 (NC_007337), which are duplicates of plasmids R100 (NC_002134), pEC-IMP (NC_012555), wild pLEW517 (NC_009132), and pJP4 (NC_005912), respectively. The final set thus consisted of 55 plasmids: 8 IncF, 7 IncH, 7 IncI, 7 IncN, 22 IncP, and 4 IncW plasmids, as shown in Table 1.

We then confirmed that the six Inc groups were clearly differentiated by their patterns of gene content (6, 64). The gene content can reflect major evolutionary phenomena such as vertical inheritance of genes, lineage-specific gene loss due to actual deletion or rapid sequence divergence, nonorthologous gene displacement, and horizontal gene transfer (27). Our all-against-all protein sequence comparisons yielded 1,318 "protein families" containing 6,097 individual proteins from the 55 plasmids (see Table S3 in the supplemental material). Only protein families that contained proteins from at least two plasmids were considered further, resulting in a final data set of 823 protein families. Based on our definition of protein families, replication initiation proteins from IncFII, IncH, IncI, IncN, IncP, and IncW belonged to different families, while

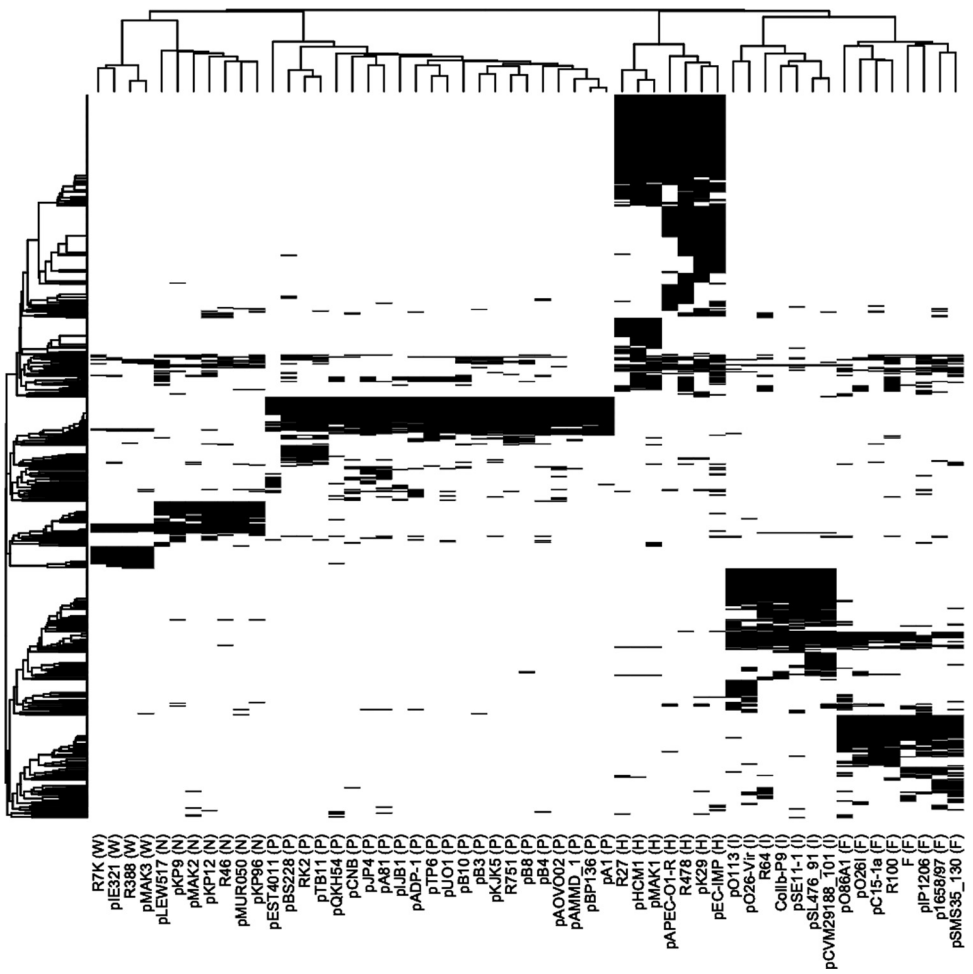

FIG. 1. Hierarchical clustering of 55 plasmids based on their dissimilarities in gene content. In the matrix, columns are plasmids, rows are protein families, and black and white denote the presence and absence of each protein family, respectively. Each character (F, H, I, N, P, and W) denotes the incompatibility group to which each plasmid belongs.

RepB and RepE (from IncFI) were homologous to RepHIA (from IncH) and RepA (from IncN), respectively. The binary data for the presence or absence of each of the protein families were subject to hierarchical clustering, and the constructed tree based upon gene content is shown in Fig. 1, where neighboring plasmids in the same cluster have similar gene content patterns. When the tree was divided into six clusters, each cluster contained the plasmids previously classified as, respectively, IncF (F and R100), IncH (R27 and R478), IncI (ColIb-P9 and R64), IncN (R46), IncP (RK2, R751, pJP4, pQKH54, and pKJK5), or IncW (R388 and R7K) by traditional incompatibility testing (Table 1). The proteins that contributed to clustering into the six groups were involved mostly in plasmid backbone functions: replication, maintenance/control, and transfer. The plasmids within the same Inc group shared many homologous genes, suggesting that they are phylogenetically closely related, while the plasmids from different Inc groups shared few homologous genes, suggesting that they are phylogenetically distantly related or have independent origins. This set of six plasmid clusters was used to compare the evolutionary host range of plasmids of the six Inc groups.

Comparison of genomic signatures among plasmids. Since the genomic signature of a plasmid tends to reflect that of the host in which it was found (67), similarities in genomic signature among plasmids may reflect similarities in evolutionary host range. In the current study, we used trinucleotide frequencies as the genomic signature of choice rather than dinucleotide frequencies, which were used in our previous study. This change was made based on the results of a performance test that compares the abilities of the two frequencies to identify known hosts (not shown). The performance test procedures were the same as those described previously (67). To visualize the similarity in genomic signature among the 55 plasmids from the 6 Inc groups, we performed multidimensional scaling to map the plasmids onto a two-dimensional space, where neighboring plasmids have similar genomic signatures (Fig. 2). Figure 2 clearly shows that plasmids belonging to the same Inc group, with similar gene content (Fig. 1), also had similar genomic signatures. Moreover, plasmids from the IncF, IncH, and IncI groups had similar signatures, while the IncN, IncP, and IncW groups had signatures that were distinctly different from each other and from those of the other three groups. Thus, while Fig. 1 shows that all these six Inc groups are

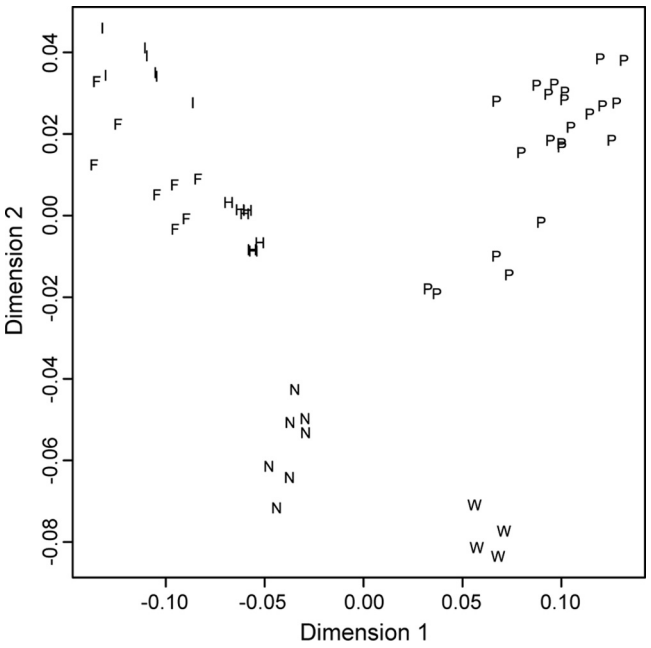

FIG. 2. A two-dimensional visualization of 55 plasmids using multidimensional scaling based on their dissimilarities in 3-mer genomic signature. Each character (F, H, I, N, P, and W) denotes the incompatibility (Inc) group to which each plasmid belongs.

phylogenetically distinct, the results in Fig. 2 suggest that the IncF, IncH, and IncI plasmids have evolved in a similar range of hosts, while the IncN, IncP, and IncW plasmids have evolved in different sets of hosts. The evolutionary host ranges are examined below.

Taxonomic range and diversity of putative evolutionary hosts of plasmids. Plasmids tend to acquire the genomic signature of their host chromosomes; therefore, the evolutionary host range of a plasmid can be inferred from the degree of similarity between the signature of the plasmid and those of sequenced bacterial chromosomes (67). Narrow-host-range plasmids are expected to have genomic signatures that are similar to only a narrow range of hosts, whereas broad-host-range plasmids are expected to show similarity with a broader range of hosts. We inferred the evolutionary host range for the 55 plasmids from the 6 Inc groups by comparing each of their signatures with each of the chromosomes of 817 prokaryotes (56 archaea and 761 bacteria). Throughout this study, a bacterial strain was proposed as a candidate evolutionary host of a given plasmid when its genomic signature was close to that of the plasmid, as indicated by high *P* values (>0.6). This relatively stringent *P* value should minimize false-positive detection of evolutionary hosts of plasmids. Indeed, all candidate evolutionary hosts were found to be members of only one phylum, the *Proteobacteria*, which is consistent with the known hosts and host ranges of these plasmids (see Table S4 in the supplemental material). Only when the threshold *P* value was lower than 0.3 were non-*Proteobacteria* detected as candidate evolutionary hosts (see Table S2 in the supplemental material). Although threshold *P* values above 0.6 resulted in more plasmids without detectable candidate hosts than now shown in Table 1, the main conclusion of this study is not altered by choosing higher or lower *P* values.

At the threshold *P* value of 0.6, no candidate evolutionary hosts were detected for any of the 4 IncW plasmids (pIE321, pMAK3, R388, and R7K) and for 2 of the 22 IncP plasmids (pA1 and R751) (Table 1). The bacterial strain most similar in genomic signature to IncW plasmids pIE321, pMAK3, and R388 was the betaproteobacterium *Dechloromonas aromatica* RCB (with low corresponding *P* values of 0.19, 0.17, and 0.12, respectively), and the most similar strain for R7K was the alphaproteobacterium *Brucella abortus* bv. 1 strain 9-941 ($P = 0.31$). The strains with signatures most similar to the IncP plasmids pA1 and R751 were the betaproteobacterium *Acidovorax* sp. JS42 ($P = 0.46$) and the gammaproteobacterium *Pseudomonas stutzeri* A1501 ($P = 0.56$), respectively. Given the low *P* values needed to detect any hosts, these six plasmids were not considered in the analyses described below.

In a first attempt to compare the evolutionary host ranges among plasmids, we inferred the kinds and diversity of candidate hosts for each plasmid by categorizing them based on their taxonomy. Table 1 shows the number of different taxa at the ranks of class, order, family, genus, and species, and Fig. 3 shows the number of candidate evolutionary hosts for all 55 plasmids, categorized by class and order. At the class level (Fig. 3A), candidate evolutionary hosts of the IncP plasmids spanned a very broad range, including up to three proteobacterial subgroups (*Alpha*-, *Beta*-, and *Gammaproteobacteria*). This is consistent with the observation that IncP plasmids are naturally found in members of at least these three classes within the phylum *Proteobacteria* (see Table S4 in the supplemental material) and that most can also transfer and replicate in representatives of these three classes in laboratory matings and microcosm studies (17, 30, 56, 72). In contrast, all candidate hosts of the IncF, IncH, IncI, and IncN plasmids belonged to the *Gammaproteobacteria* only. At the order level, Fig. 3B also clearly shows that candidate evolutionary hosts for the IncP group encompassed a much wider taxonomic range than that for the IncF, IncH, IncI, and IncN groups (mostly *Enterobacteriales*), with up to seven orders for some IncP plasmids. This result for the IncF and IncI plasmids is in agreement with experimental evidence that their host range is limited to members of the *Enterobacteriales*, while those for the IncH and IncN plasmids warrant further discussion (see Discussion) (13). In conclusion, the results suggest that IncP plasmids have evolved in the widest taxonomic range, while IncF, IncH, IncI, and IncN plasmids have resided in a narrower taxonomic range.

Second, we inferred plasmid host range based on genetic distance between putative evolutionary hosts without using taxonomic information. "Taxonomic richness" has drawbacks, because it does not take into account distances between bacterial strains and also can be influenced by the number of strains detected as candidate evolutionary hosts, which in turn is biased by the genome sequences available. For example, as shown in Table 1, the numbers of classes, orders, families, genera, species, and strains for IncP plasmid pQKH54 were 2, 3, 3, 3, 6, and 11, respectively, while those for IncN plasmid R46 were 1, 1, 1, 13, 17, and 55. Thus, while pQKH54 showed higher taxonomic richness values at the class, order, and family levels, R46 showed higher taxonomic richness values at the genus, species, and strain levels. Moreover, several bacterial

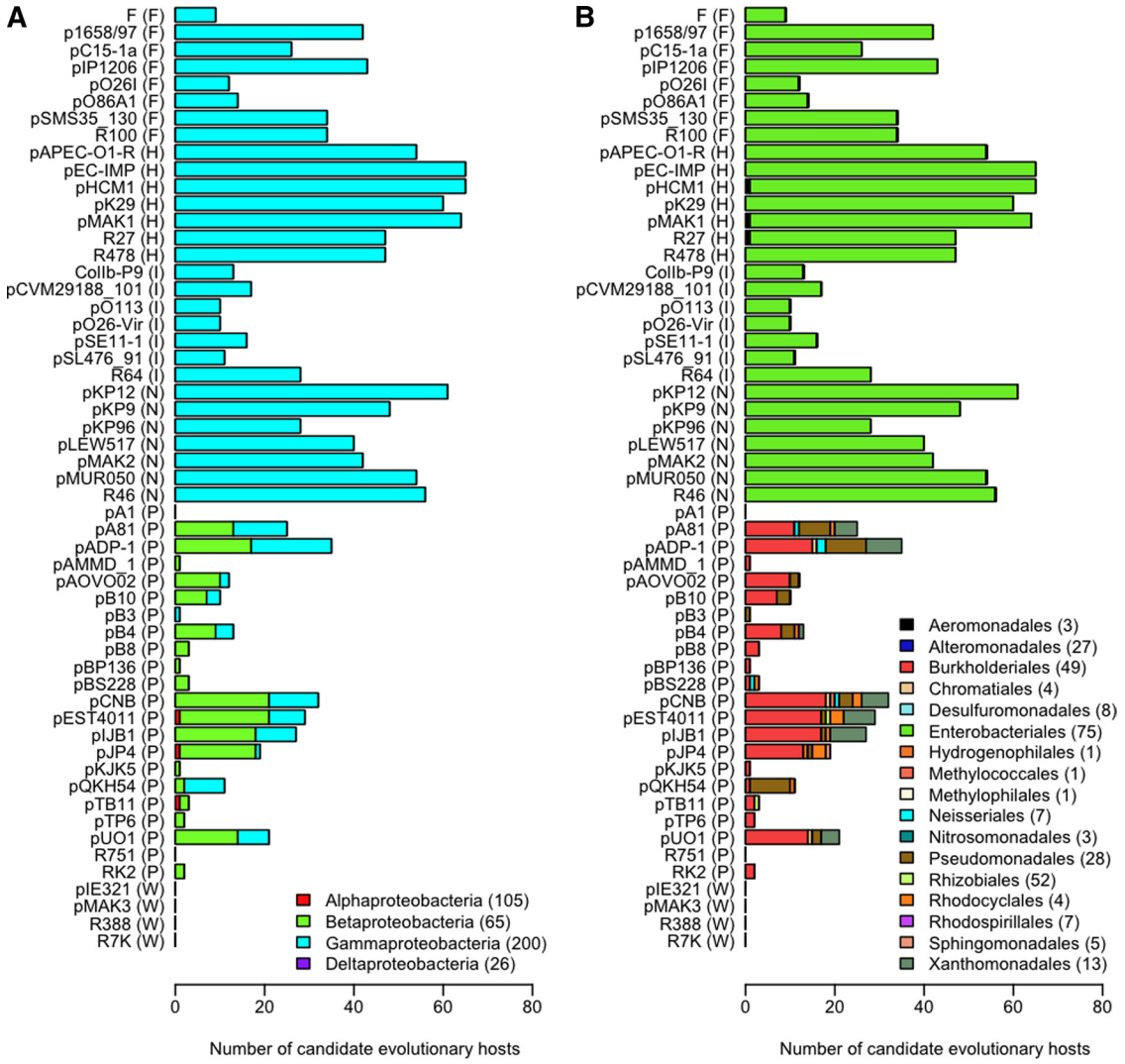

FIG. 3. Bar plot showing the number of candidate evolutionary hosts for each plasmid. Different colors indicate different taxonomic groups at the level of class (A) and order (B). The number of strains belonging to each taxon is given in parentheses. Each character (F, H, I, N, P, and W) denotes the incompatibility (Inc) group to which each plasmid belongs.

strains are still being reclassified into new species and even new genera. Therefore, a method for inferring evolutionary host diversity that does not require a species or any other taxonomic information was conducted.

We quantified the diversity of candidate evolutionary hosts for each plasmid by using the mean genetic distance (*D*mean) between all pairs of candidate hosts (76). Distances were measured in two ways: on the basis of the plasmid hosts' dissimilarity in 16S rRNA gene sequence and dissimilarity in 3-mer genomic signature. As a point of reference, the *D*mean values between all 817 strains used in the analyses based on these two features were 0.281 and 0.255, respectively. The *D*mean values for individual plasmids are shown in Table 1 and summarized per Inc group in Fig. 4 as box-and-whisker plots. The diversity of plasmid hosts based on the distances between their 16S rRNA gene sequences was highest for the IncP plasmids (median *D*mean value of 0.128), followed by those of the IncH (0.034), IncN (0.029), IncF (0.013), and IncI (0.008) groups

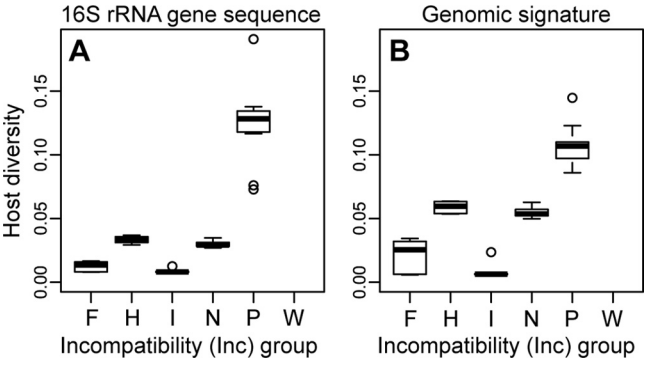

FIG. 4. Box-and-whisker plots summarizing the distributions of diversity among candidate evolutionary hosts for plasmids from six incompatibility (Inc) groups (F, H, I, N, P, and W). The host diversity for each plasmid was quantified by the mean distance (*D*mean) between all host pairs based on their dissimilarity in 16S rRNA gene sequence (A) and 3-mer genomic signature (B).

(the difference was statistically significant based on a Kruskal-Wallis rank sum test; P value of $<10^{-6}$). Similar results were obtained when using alignments after removing any alignment position where any of the sequences showed a gap (data not shown). The diversity of plasmid hosts based on the distances between their 3-mer genomic signatures was again highest for the IncP plasmids (median Dmean value of 0.107), followed by those of the IncH (0.060), IncN (0.054), IncF (0.025), and IncI (0.006) plasmids (Kruskal-Wallis rank sum test; P value of $<10^{-6}$). These findings confirm that the range of candidate evolutionary hosts for the IncP plasmids is wider than the ranges of the IncF, IncH, IncI, and IncN plasmids, suggesting that IncP plasmids have been much more promiscuous over evolutionary time than their IncF, IncH, IncI, and IncN counterparts. The only caveat is for the IncW group, as the candidate host diversity could not be tested due to lack of sufficient candidate hosts.

Finally, we assessed the evolutionary host range of plasmids from additional groups of interest with multiple sequenced representatives: 10 IncA/C, 3 IncL/M, 4 IncP-9, 7 IncQ, 2 IncU, 4 PromA, and 7 Ri/Ti plasmids (see Table S4 in the supplemental material). At the threshold P value of 0.6, no or less than three hosts were detected for 4 of the 10 IncA/C plasmids tested, all 7 IncQ plasmids, both IncU plasmids, and 3 of the 4 PromA plasmids. This result was similar to that for all 4 IncW plasmids and 8 IncP plasmids reported above (Table 1 and Fig. 3) and requires further investigation. All candidate evolutionary hosts detected for the remaining plasmids were members of the *Proteobacteria* (Table 1) (see Fig. S1 in the supplemental material): *Beta-*, *Gamma-*, or *Deltaproteobacteria* for the IncA/C plasmids, *Beta-* and *Gammaproteobacteria* for the IncP-9 plasmids, and only *Gammaproteobacteria* for the IncL/M plasmids. As expected, the commonly predicted hosts of the Ri/Ti plasmids were the *Rhizobiales* (see Fig. S1 in the supplemental material). The predicted evolutionary host diversity for the IncA/C, IncP-9, and Ri/Ti plasmids, as expressed by median Dmean values (based on 16S rRNA gene sequences), was higher than that for the IncF, IncH, IncI, IncN, and IncL/M plasmids. Similar results were obtained when different threshold P values were used (see Fig. S2 in the supplemental material). Together, these results suggest that the IncA/C, IncP, IncP-9, and Ri/Ti plasmids have broader evolutionary host ranges than the IncF, IncH, IncI, IncL/M, and IncN plasmids.

DISCUSSION

In this study, we used similarities in genomic signatures between 55 plasmids belonging to 6 well-studied incompatibility groups and all available prokaryotic chromosome sequences to infer the plasmids' evolutionary host range, i.e., the range of hosts in which these plasmids may have resided over evolutionary time. Generally, IncF, IncH, and IncI group plasmids are considered to have a narrow host range, while IncN, IncP, and IncW group plasmids have a broad host range (13, 49, 62). The results of a genomic signature comparison suggest that the broad-host-range IncP plasmids indeed have a broad evolutionary host range and that the narrow-host-range IncF and IncI plasmids have the narrowest (Fig. 4). This indicates that a plasmid's genomic signature can provide information about its promiscuity. The inferred intermediate evolutionary host range for the IncH and IncN plasmids, and the absence of putative evolutionary hosts for IncW plasmids, can also in part be explained after a more detailed analysis of previous studies, as described below. We then inferred the evolutionary host range of self-transmissible plasmids whose host ranges have been studied in less detail (IncA/C, IncL/M, IncP-9 and IncU, PromA and Ri/Ti) and the non-self-transmissible IncQ plasmids. Below, we discuss our findings on the predicted evolutionary host range for each of these plasmid groups in light of knowledge gained from previous empirical studies on their replication and transfer range. This is the first report of a thoroughly validated genomic tool that allows the inference of the putative host range of plasmids or other mobile elements solely on the basis of their DNA sequence.

We showed that plasmids that belong to the same Inc group with similar gene content (Fig. 1) also have similar genomic signatures (Fig. 2). It is expected that closely related plasmids within the same Inc group that share a recent common ancestor are similar in gene content and genomic signature. However, after continued divergence, these plasmids could possibly acquire different genomic signatures. Since this is not what we observed, and given our current understanding of the diversity in plasmid replication machineries and their complex interactions with several host factors (18, 50, 73), we posit the following. Because plasmids of the same Inc group have similar replication machineries, they have been restricted to a similar range of hosts; therefore, their genomic signatures have been driven to be compatible with that of these hosts. It is no surprise then that their signatures remained similar even after the plasmids diverged from a common ancestor.

The intermediate evolutionary host range for the IncH plasmids may seem contradictory to the common assumption that IncH plasmids have a narrow host range. However, a closer look at the experimental studies shows that subgroup IncHI plasmids are thermosensitive for conjugative transfer; i.e., transfer efficiency is optimal at temperatures below 30°C (70). For example, it was shown that at 24°C, the host range of most IncHI plasmids was equivalent to that of broad-host-range plasmids IncN, IncP, and IncW, and that at 14°C, IncHI plasmid R478 showed the broadest host range and transfer proficiency of any plasmid tested (47). Thus, while the host range of IncH seems narrow at common laboratory temperatures of 30 to 37°C, at lower temperatures in many natural environments, IncH plasmids may well transfer between and reside in a moderately wide range of hosts.

At first sight, our finding that IncN plasmids have an intermediate evolutionary host range (*Enterobacteriales* only) does not support the commonly found statement that IncN plasmids are broad-host-range plasmids (62) and some findings that corroborate that statement. For example, IncN plasmids have been transferred to the alphaproteobacterium *Caulobacter crescentus* (22) and even the deltaproteobacterium *Myxococcus* sp. (55). However, there are also suggestions that IncN plasmids have a more limited host range. First, IncN plasmids are most often found in *Enterobacteriales*, such as *Escherichia*, *Klebsiella*, *Proteus*, *Providencia*, *Salmonella*, *Shigella*, and *Yersinia* species (62). Second, compared to IncP plasmid RK2, IncN plasmids were found either to transfer inefficiently or to be unstable in nonenteric bacteria (69). This observation of

limited host range was also confirmed in an independent study that compared the IncN plasmid pCU1 with the IncP plasmid RK2 (40), as well as in a soil microcosm study (56). Thus, whereas IncN plasmids can potentially transfer and replicate in a broad range of hosts, rigorous studies show that this range is narrower than for IncP plasmids. These findings thus support our inference that the evolutionary host range of IncN plasmids is wider than that of IncF and IncI plasmids but narrower than that of IncP plasmids.

IncW plasmids are also considered to have a broad host range, because they have been found in a wide variety of bacteria, including *Alpha-*, *Beta-*, *Gamma-*, *Deltaproteobacteria*, and *Bacteroidetes* (24). Interestingly, at the threshold *P* value of 0.6, not a single candidate evolutionary host was detected from among 817 strains for any of the four IncW plasmids (Fig. 3 and Table 1). The IncA/C, IncP, IncQ, IncU, and PromA groups also contained such plasmids whose genomic signatures were not similar to those of any of the chromosomes. There are at least two possible explanations for not detecting candidate evolutionary hosts for these broad-host-range plasmids. The first is that the evolutionary hosts of these plasmids have not yet been isolated and sequenced. This hypothesis cannot be entirely excluded, since a large proportion of prokaryotes have not yet been completely sequenced. While the complete genome sequences of the strains in which the IncW plasmids were found (Table 1) (see Table S1 in the supplemental material) have not been determined, genome sequences of related strains of the same species are available, yet do not show high signature similarity with these plasmids. For example, the *P* values for plasmid pIE321 and the genomes of 13 *Salmonella enterica* subsp. *enterica* strains and those for plasmid R388 and the 22 *Escherichia coli* strains were lower than 0.01, indicating significantly different genomic signatures. This suggests that the known hosts listed in Table 1 have only recently acquired these plasmids. Second, these plasmids have likely horizontally transferred between multiple diverse hosts. Due to this promiscuous history, there may not have been sufficient time for plasmid genome amelioration to occur in any host, or their genomic signatures reflect a mixture of diverse host signatures. According to the second hypothesis, plasmids like those of the IncW group for which no candidate evolutionary hosts are detected would have a broader evolutionary host range than plasmids for which diverse hosts can be detected, such as the IncP plasmids. Future experimental and genomic analyses should test this hypothesis.

When the signature method was further tested with a second set of plasmids whose host ranges have been studied in less detail, both expected and surprising results were obtained. The candidate evolutionary hosts for the Ri/Ti plasmids were mostly *Rhizobiales* (see Fig. S1 in the supplemental material), which is in agreement with their known hosts (Table 1; also, see Table S4 in the supplemental material). The lack of signature similarity with any host for most IncQ, IncU and PromA plasmids, suggesting a high level of promiscuity, is consistent with empirical studies that have shown wide host ranges for plasmids from these groups, including at least three proteobacterial classes (41, 74). The IncQ plasmids, the only non-self-transmissible plasmids included in the study, can also replicate in Gram-positive bacteria (58). In contrast to an empirical study that showed replication of IncL/M plasmid pCTX-M3 in members of three proteobacterial classes (*Alpha-*, *Beta-*, and *Gammaproteobacteria*) (52), we predicted the plasmids from this group to have an intermediate evolutionary host range (*Enterobacteriales* only), similar to that of the IncN plasmid (Table 1) (see Fig. S1 and S2 in the supplemental material). IncL/M and IncN plasmids also have similar genomic signatures (data not shown), suggesting that they have evolved in similar ranges of hosts. In addition, our method predicted that the evolutionary host range of the IncA/C and IncP-9 plasmids is as broad as or broader than that of the IncP plasmids (see Fig. S1 and S2 in the supplemental material), whereas so far as we know, representative plasmids RA1 (46) and pWW0 (31) have been shown to replicate only in *Gammaproteobacteria*. In conclusion, the predicted evolutionary host ranges were as expected for the IncQ, IncU, Ri/Ti, and PromA plasmids, narrower than expected based on limited experimental data for the IncL/M plasmids, and broader than expected for the IncA/C and IncP-9 plasmids. The discrepancy should be examined further in future studies.

All analyses in this study were done using the entire plasmid genomes. We are aware of the mosaic nature of plasmids, which contain discrete DNA regions; i.e., those vertically inherited from a common ancestral plasmid and those recently acquired by horizontal transfer from the host chromosomes or other mobile elements. The vertically inherited plasmid regions should be similar in genomic signature to those of their long-term evolutionary hosts due to genome amelioration. Plasmid fragments acquired from the host chromosomes may also result in similar signatures between plasmids and their hosts even after short-term residence. To address the question of chromosomal insertions in plasmids, we did a preliminary test in which plasmid sequences that were detected in any of the 817 chromosomes were eliminated from the plasmids before analysis. The results did not change our main conclusion, i.e., that the evolutionary host range is broad for IncP plasmids, narrow for IncF and IncI plasmids, and intermediate for IncH and IncN plasmids (data not shown). However, acquisition by a plasmid of DNA from other mobile elements with very different genomic signatures, for example through recombination or transposition between coresiding plasmids with different host ranges, may result in a combination of very different genomic signatures. When the history of plasmid fragments is that different, our test may fail to detect the evolutionary plasmid hosts. The broader than expected evolutionary host range for the IncA/C and IncP-9 plasmids could be due in part to such gene acquisition. Future studies will need to unravel the effect of horizontal gene transfer between plasmids on plasmid signatures and the signature-based prediction of their evolutionary host range.

In summary, this study establishes that the host range of plasmids can be inferred from their genomic signatures. The lack of hosts with genomic signatures similar to those of the IncW and several other broad-host-range plasmids is intriguing and requires further investigation. Genome sequence analysis of more bacterial chromosomes and plasmids is required to minimize sampling bias and maximize phylogenetic coverage (78). The discrepancies between inferred evolutionary host ranges and empirically determined replication ranges for some plasmid groups require further investigation. While the plasmid data set used in this present study may be limited and

biased, future experimental and genomic studies will improve our understanding of the evolutionary host range of plasmids. This genomic tool to assess plasmid host range will thus provide insight into the promiscuity and potential reservoirs of plasmids and other mobile genetic elements in the horizontal gene pool (65). This is not only of interest to the fields of plasmid biology and bacterial evolution but also of medical relevance, given that many of these plasmids threaten our ability to adequately combat infectious diseases (44).

ACKNOWLEDGMENTS

We thank Max Mergeay and members of IBEST at the University of Idaho and of the Institute for Advanced Biosciences at Keio University for useful discussions. We are also grateful to Kazuharu Arakawa for his advice on G-language genome analysis environment, to Jan Mrázek for his advice on calculating genomic signature, to Anthony Haines for his advice on plasmid RK2, and to Jacob Pierson and Ursel Schütte for their advice on measuring bacterial diversity.

This work was supported by the National Science Foundation (grants EF-0627988 and DBI-0939454). H. Yano, C. Brown, and E. Top were also supported by a COBRE grant from the National Institutes of Health, National Center for Research Resources (P20RR16448). The Bioinformatics facilities were supported by the same COBRE grant as well as by INBRE from the same NIH Center (P20RR016454).

REFERENCES

1. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
2. **Arakawa, K., K. Mori, K. Ikeda, T. Matsuzaki, Y. Kobayashi, and M. Tomita.** 2003. G-language genome analysis environment: a workbench for nucleotide sequence data mining. Bioinformatics **19:**305–306.
3. **Asano, K., and K. Mizobuchi.** 1998. An RNA pseudoknot as the molecular switch for translation of the *repZ* gene encoding the replication initiator of IncIalpha plasmid ColIb-P9. J. Biol. Chem. **273:**11815–11825.
4. **Bahl, M. I., L. H. Hansen, A. Goesmann, and S. J. Sorensen.** 2007. The multiple antibiotic resistance IncP-1 plasmid pKJK5 isolated from a soil environment is phylogenetically divergent from members of the previously established alpha, beta and delta sub-groups. Plasmid **58:**31–43.
5. **Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers.** 2009. GenBank. Nucleic Acids Res. **37:**D26–D31.
6. **Brilli, M., A. Mengoni, M. Fondi, M. Bazzicalupo, P. Lio, and R. Fani.** 2008. Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network. BMC Bioinformatics **9:**551.
7. **Campbell, A., J. Mrazek, and S. Karlin.** 1999. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc. Natl. Acad. Sci. U. S. A. **96:**9184–9189.
8. **Chiu, C. M., and C. M. Thomas.** 2004. Evidence for past integration of IncP-1 plasmids into bacterial chromosomes. FEMS Microbiol. Lett. **241:** 163–169.
9. **Choi, I. G., and S. H. Kim.** 2007. Global extent of horizontal gene transfer. Proc. Natl. Acad. Sci. U. S. A. **104:**4489–4494.
10. **Coetzee, J. N., N. Datta, and R. W. Hedges.** 1972. R factors from *Proteus rettgeri*. J. Gen. Microbiol. **72:**543–552.
11. **Couturier, M., F. Bex, P. L. Bergquist, and W. K. Maas.** 1988. Identification and classification of bacterial plasmids. Microbiol. Rev. **52:**375–395.
12. **Datta, N., and R. W. Hedges.** 1971. Compatibility groups among *fi*$^-$ R factors. Nature **234:**222–223.
13. **Datta, N., and R. W. Hedges.** 1972. Host ranges of R factors. J. Gen. Microbiol. **70:**453–460.
14. **Datta, N., and R. W. Hedges.** 1972. Trimethoprim resistance conferred by W plasmids in Enterobacteriaceae. J. Gen. Microbiol. **72:**349–355.
15. **Datta, N., R. W. Hedges, E. J. Shaw, R. B. Sykes, and M. H. Richmond.** 1971. Properties of an R factor from *Pseudomonas aeruginosa*. J. Bacteriol. **108:** 1244–1249.
16. **De Gelder, L., J. M. Ponciano, P. Joyce, and E. M. Top.** 2007. Stability of a promiscuous plasmid in different hosts: no guarantee for a long-term relationship. Microbiology **153:**452–463.
17. **De Gelder, L., F. P. Vandecasteele, C. J. Brown, L. J. Forney, and E. M. Top.** 2005. Plasmid donor affects host range of promiscuous IncP-1β plasmid pB10 in an activated-sludge microbial community. Appl. Environ. Microbiol. **71:**5309–5317.
18. **del Solar, G., J. C. Alonso, M. Espinosa, and R. Diaz-Orejas.** 1996. Broad-host-range plasmid replication: an open question. Mol. Microbiol. **21:**661–666.
19. **Delver, E. P., and A. A. Belogurov.** 1997. Organization of the leading region of IncN plasmid pKM101 (R46): a regulation controlled by CUP sequence elements. J. Mol. Biol. **271:**13–30.
20. **Don, R. H., and J. M. Pemberton.** 1981. Properties of six pesticide degradation plasmids isolated from *Alcaligenes paradoxus* and *Alcaligenes eutrophus*. J. Bacteriol. **145:**681–686.
21. **Edgar, R. C.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:**1792–1797.
22. **Ely, B.** 1979. Transfer of drug resistance factors to the dimorphic bacterium *Caulobacter crescentus*. Genetics **91:**371–380.
23. **Fang, F. C., and D. R. Helinski.** 1991. Broad-host-range properties of plasmid RK2: importance of overlapping genes encoding the plasmid replication initiation protein TrfA. J. Bacteriol. **173:**5861–5868.
24. **Fernandez-Lopez, R., M. P. Garcillan-Barcia, C. Revilla, M. Lazaro, L. Vielva, and F. de la Cruz.** 2006. Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution. FEMS Microbiol. Rev. **30:**942–966.
25. **Fielding, A. H.** 2007. Cluster and classification techniques for the biosciences. Cambridge University Press, New York, NY.
26. **Frost, L. S., R. Leplae, A. O. Summers, and A. Toussaint.** 2005. Mobile genetic elements: the agents of open source evolution. Nat. Rev. Microbiol. **3:**722–732.
27. **Galperin, M. Y., and E. V. Koonin.** 2000. Who's your neighbor? New computational approaches for functional genomics. Nat. Biotechnol. **18:**609–613.
28. **Garcillan-Barcia, M. P., M. V. Francia, and F. de la Cruz.** 2009. The diversity of conjugative relaxases and its application in plasmid classification. FEMS Microbiol. Rev. **33:**657–687.
29. **Gardner, R., J. McAnulty, E. Feher, and D. Lane.** 1985. Location of *rep* and *inc* sequences in the F secondary replicon. Plasmid **13:**145–148.
30. **Goris, J., W. Dejonghe, E. Falsen, E. De Clerck, B. Geeraerts, A. Willems, E. M. Top, P. Vandamme, and P. De Vos.** 2002. Diversity of transconjugants that acquired plasmid pJP4 or pEMT1 after inoculation of a donor strain in the A- and B-horizon of an agricultural soil and description of *Burkholderia hospita* sp. nov. and *Burkholderia terricola* sp. nov. Syst. Appl. Microbiol. **25:**340–352.
31. **Greated, A., L. Lambertsen, P. A. Williams, and C. M. Thomas.** 2002. Complete sequence of the IncP-9 TOL plasmid pWW0 from *Pseudomonas putida*. Environ. Microbiol. **4:**856–871.
32. **Guiney, D. G., P. Hasegawa, and C. E. Davis.** 1984. Plasmid transfer from *Escherichia coli* to *Bacteroides fragilis*: differential expression of antibiotic resistance phenotypes. Proc. Natl. Acad. Sci. U. S. A. **81:**7203–7206.
33. **Haines, A. S., P. Akhtar, E. R. Stephens, K. Jones, C. M. Thomas, C. D. Perkins, J. R. Williams, M. J. Day, and J. C. Fry.** 2006. Plasmids from freshwater environments capable of IncQ retrotransfer are diverse and include pQKH54, a new IncP-1 subgroup archetype. Microbiology **152:**2689–2701.
34. **Hedges, R. W., and N. Datta.** 1971. *fi*$^-$ R factors giving chloramphenicol resistance. Nature **234:**220–221.
35. **Hedges, R. W., and N. Datta.** 1972. R124, an *fi* R factor of a new compatibility class. J. Gen. Microbiol. **71:**403–405.
36. **Ingram, L. C., M. H. Richmond, and R. B. Sykes.** 1973. Molecular characterization of the R factors implicated in the carbenicillin resistance of a sequence of *Pseudomonas aeruginosa* strains isolated from burns. Antimicrob. Agents Chemother. **3:**279–288.
37. **Jacob, A. E., J. A. Shapiro, L. Yamamoto, D. I. Smith, S. N. Cohen, and D. Berg.** 1977. Plasmids studied in *Escherichia coli* and other enteric bacteria, p. 607–638. *In* A. I. Bukhari, J. A. Shapiro, and S. L. Adhya (ed.), DNA insertion elements, plasmids, and episomes. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
38. **Jiang, T., Y. N. Min, W. Liu, D. D. Womble, and R. H. Rownd.** 1993. Insertion and deletion mutations in the *repA4* region of the IncFII plasmid NR1 cause unstable inheritance. J. Bacteriol. **175:**5350–5358.
39. **Jobanputra, R. S., and N. Datta.** 1974. Trimethoprim R factors in enterobacteria from clinical specimens. J. Med. Microbiol. **7:**169–177.
40. **Krishnan, B. R., and V. N. Iyer.** 1988. Host ranges of the IncN group plasmid pCU1 and its minireplicon in gram-negative purple bacteria. Appl. Environ. Microbiol. **54:**2273–2276.
41. **Kulinska, A., M. Czeredys, F. Hayes, and G. Jagura-Burdzy.** 2008. Genomic and functional characterization of the modular broad-host-range RA3 plasmid, the archetype of the IncU group. Appl. Environ. Microbiol. **74:**4119–4132.
42. **Lavigne, J. P., A. C. Vergunst, G. Bourg, and D. O'Callaghan.** 2005. The IncP island in the genome of *Brucella suis* 1330 was acquired by site-specific integration. Infect. Immun. **73:**7779–7783.
43. **Lawrence, J. G., and H. Ochman.** 1997. Amelioration of bacterial genomes: rates of change and exchange. J. Mol. Evol. **44:**383–397.
44. **Levy, S. B., and B. Marshall.** 2004. Antibacterial resistance worldwide: causes, challenges and responses. Nat. Med. **10:**S122–S129.
45. **Llanes, C., P. Gabant, M. Couturier, L. Bayer, and P. Plesiat.** 1996. Molecular analysis of the replication elements of the broad-host-range RepA/C replicon. Plasmid **36:**26–35.
46. **Llanes, C., P. Gabant, M. Couturier, and Y. Michel-Briand.** 1994. Cloning

and characterization of the Inc A/C plasmid RA1 replicon. J. Bacteriol. **176:**3403–3407.
47. **Maher, D., and D. E. Taylor.** 1993. Host range and transfer efficiency of incompatibility group HI plasmids. Can. J. Microbiol. **39:**581–587.
48. **Masson, L., and D. S. Ray.** 1988. Mechanism of autonomous control of the *Escherichia coli* F plasmid: purification and characterization of the *repE* gene product. Nucleic Acids Res. **16:**413–424.
49. **Mazodier, P., and J. Davies.** 1991. Gene transfer between distantly related bacteria. Annu. Rev. Genet. **25:**147–171.
50. **Meyer, R.** 2009. Replication and conjugative mobilization of broad host-range IncQ plasmids. Plasmid **62:**57–70.
51. **Meynell, E., and N. Datta.** 1966. The nature and incidence of conjugation factors in *Escherichia coli*. Genet. Res. **7:**141–148.
52. **Mierzejewska, J., A. Kulinska, and G. Jagura-Burdzy.** 2007. Functional analysis of replication and stability regions of broad-host-range conjugative plasmid CTX-M3 from the IncL/M incompatibility group. Plasmid **57:**95–107.
53. **Novick, R. P.** 1987. Plasmid incompatibility. Microbiol. Rev. **51:**381–395.
54. **Pansegrau, W., E. Lanka, P. T. Barth, D. H. Figurski, D. G. Guiney, D. Haas, D. R. Helinski, H. Schwab, V. A. Stanisich, and C. M. Thomas.** 1994. Complete nucleotide sequence of Birmingham IncP alpha plasmids: compilation and comparative analysis. J. Mol. Biol. **239:**623–663.
55. **Parish, J. H.** 1975. Transfer of drug resistance to *Myxococcus* from bacteria carrying drug-resistance factors. J. Gen. Microbiol. **87:**198–210.
56. **Pukall, R., H. Tschape, and K. Smalla.** 1996. Monitoring the spread of broad host and narrow host range plasmids in soil microcosms. FEMS Microbiol. Ecol. **20:**53–66.
57. **R Development Core Team.** 2009. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
58. **Rawlings, D. E., and E. Tietze.** 2001. Comparative biology of IncQ and IncQ-like plasmids. Microbiol. Mol. Biol. Rev. **65:**481–496.
59. **Scherzinger, E., V. Haring, R. Lurz, and S. Otto.** 1991. Plasmid RSF1010 DNA replication *in vitro* promoted by purified RSF1010 RepA, RepB and RepC proteins. Nucleic Acids Res. **19:**1203–1211.
60. **Sevastsyanovich, Y. R., R. Krasowiak, L. E. Bingle, A. S. Haines, S. L. Sokolov, I. A. Kosheleva, A. A. Leuchuk, M. A. Titok, K. Smalla, and C. M. Thomas.** 2008. Diversity of IncP-9 plasmids of *Pseudomonas*. Microbiology **154:**2929–2941.
61. **Sevastsyanovich, Y. R., M. A. Titok, R. Krasowiak, L. E. Bingle, and C. M. Thomas.** 2005. Ability of IncP-9 plasmid pM3 to replicate in *Escherichia coli* is dependent on both *rep* and *par* functions. Mol. Microbiol. **57:**819–833.
62. **Shapiro, J. A.** 1977. Bacterial plasmids (appendix B), p. 601–704. *In* A. I. Bukhari, J. A. Shapiro, and S. L. Adhya (ed.), DNA insertion elements, plasmids, and episomes. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
63. **Sherburne, C. K., T. D. Lawley, M. W. Gilmour, F. R. Blattner, V. Burland, E. Grotbeck, D. J. Rose, and D. E. Taylor.** 2000. The complete DNA sequence and analysis of R27, a large IncHI plasmid from *Salmonella typhi* that is temperature sensitive for transfer. Nucleic Acids Res. **28:**2177–2186.
64. **Snel, B., P. Bork, and M. A. Huynen.** 1999. Genome phylogeny based on gene content. Nat. Genet. **21:**108–110.
65. **Sommer, M. O., G. Dantas, and G. M. Church.** 2009. Functional characterization of the antibiotic resistance reservoir in the human microflora. Science **325:**1128–1131.
66. **Summers, A. O.** 2006. Genetic linkage and horizontal gene transfer, the roots of the antibiotic multi-resistance problem. Anim. Biotechnol. **17:**125–135.
67. **Suzuki, H., M. Sota, C. J. Brown, and E. M. Top.** 2008. Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes. Nucleic Acids Res. **36:**e147.
68. **Suzuki, K., K. Tanaka, S. Yamamoto, K. Kiyokawa, K. Moriguchi, and K. Yoshida.** 2009. Ti and Ri plasmids, p. 133–147. *In* E. Schwartz (ed.), Microbial megaplasmids. Springer, Berlin, Germany.
69. **Tardif, G., and R. B. Grant.** 1980. Characterisation of the host range of the N-plasmids, p. 351–359. *In* C. Stuttard and K. R. Rozee (ed.), Plasmids and transposons: environmental and maintenance mechanism. Academic Press, New York, NY.
70. **Taylor, D. E.** 2009. Thermosensitive nature of IncHI1 plasmid transfer. Antimicrob. Agents Chemother. **53:**2703.
71. **Taylor, D. E., and R. B. Grant.** 1977. R plasmids of the S incompatibility group belong to the H2 incompatibility group. Antimicrob. Agents Chemother. **12:**431–434.
72. **Thomas, C. M., and C. A. Smith.** 1987. Incompatibility group P plasmids: genetics, evolution, and use in genetic manipulation. Annu. Rev. Microbiol. **41:**77–101.
73. **Toukdarian, A.** 2004. Plasmid strategies for broad-host-range replication in gram-negative bacteria, p. 259–270. *In* B. E. Funnell and and G. J. Phillips (ed.), Plasmid biology. ASM Press, Washington, DC.
74. **Van der Auwera, G. A., J. E. Krol, H. Suzuki, B. Foster, R. Van Houdt, C. J. Brown, M. Mergeay, and E. M. Top.** 2009. Plasmids captured in *C. metallidurans* CH34: defining the PromA family of broad-host-range plasmids. Antonie Van Leeuwenhoek **96:**193–204.
75. **van Dongen, S.** 2000. Graph clustering by flow simulation. Ph.D. thesis. University of Utrecht, Utrecht, Netherlands.
76. **Watve, M. G., and R. M. Gangal.** 1996. Problems in measuring bacterial diversity and a possible solution. Appl. Environ. Microbiol. **62:**4299–4301.
77. **Womble, D. D., and R. H. Rownd.** 1988. Genetic and physical map of plasmid NR1: comparison with other IncFII antibiotic resistance plasmids. Microbiol. Rev. **52:**433–451.
78. **Wu, D., P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. D'Haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J. F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H. P. Klenk, and J. A. Eisen.** 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature **462:**1056–1060.
79. **Wu, R. P., D. D. Womble, and R. H. Rownd.** 1985. Incompatibility mutants of IncFII plasmid NR1 and their effect on replication control. J. Bacteriol. **163:**973–982.